Experimental design for microbiome studies

Mihai Pop
University of Maryland
College Park, MD

Experimental design

- Samples
- Measurements
- Analyses

Which one do you figure out first?

How do you make the decision?

The samples

- Cross-sectional versus longitudinal?
- If longitudinal, how frequently?
- N = ?

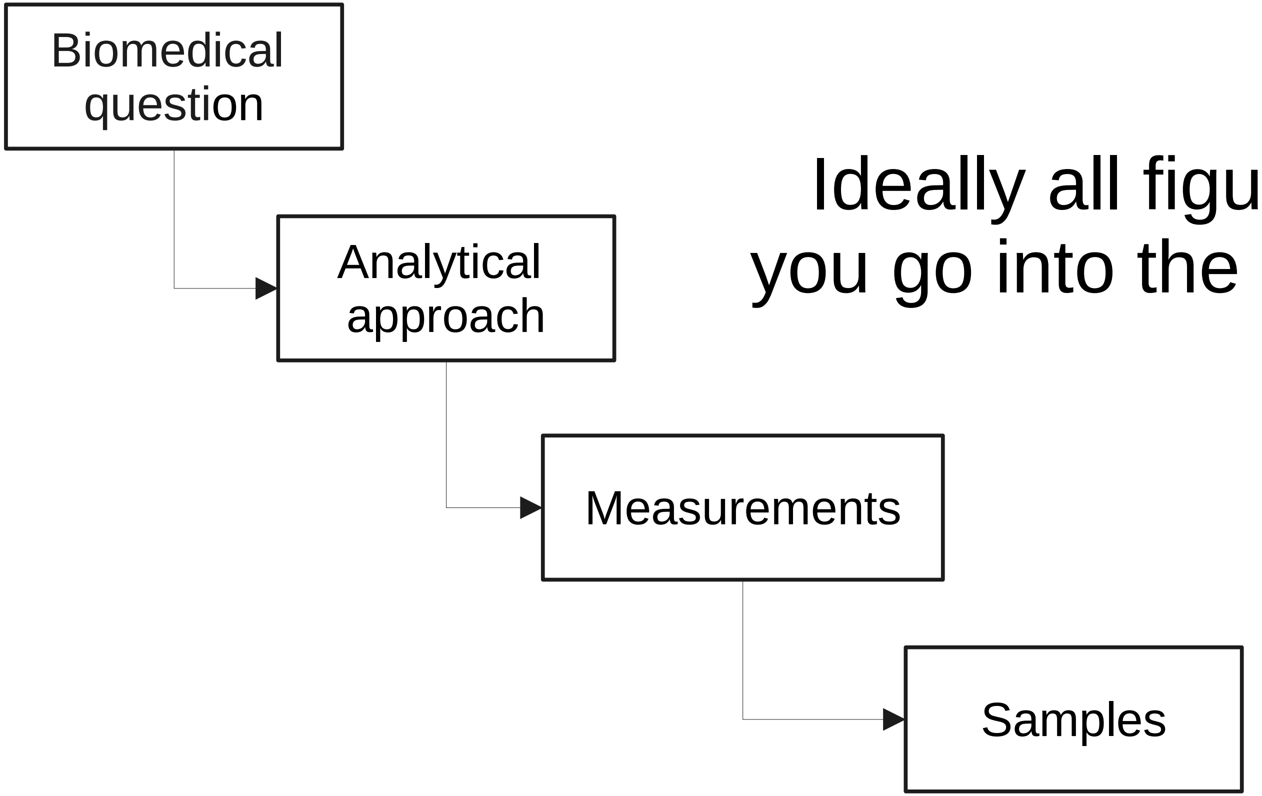
The measurements

- 16S or metagenomics?
- Transcriptomics? Metabolomics?
- Single cell data?
- qPCR?
- Short read vs. long read?
- What metadata ?

The analysis

- What k-mer size?
- What assembly/clustering/binning approach?
- What statistical test/software?
- etc.

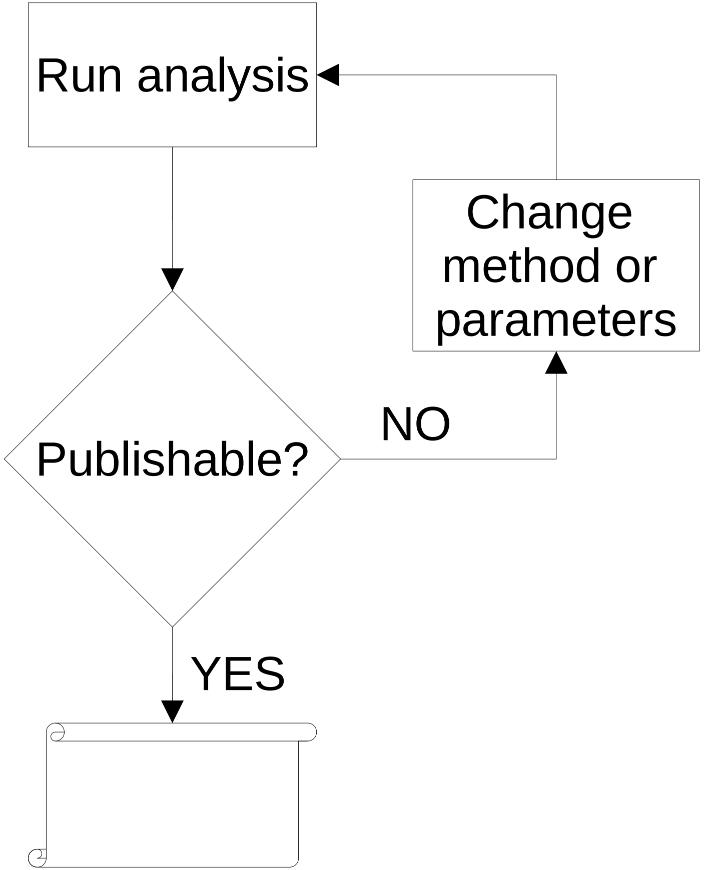
Backwards design



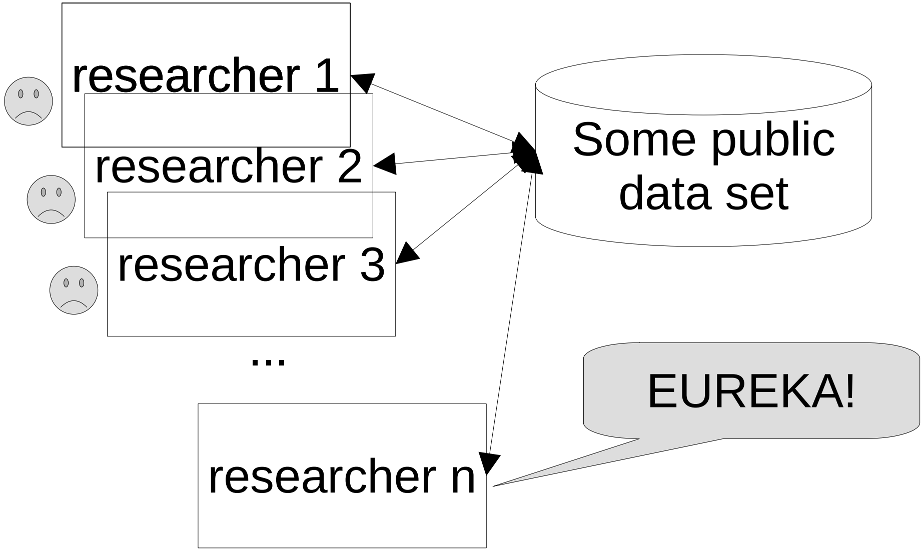
Ideally all figured out before you go into the clinic / field / lab!

* clinical studies tend to do this fairly well

Beware multiple testing



IMPORTANT: Problem exists even if you are not the one re-running analyses (relevant when using public data)!



Christie Aschwanden. *Science isn't broken*. https://fivethirtyeight.com/features/science-isnt-broken/ (with interactive tool)

Aside: the declining value of data

- Differential privacy "math" can quantify how much you "overfit" (what information is leaked from the data)
- Each access (by anyone) decreases the value of the data set (for machine learning/inference)

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2017. *Guilt-free data reuse*. Commun. ACM 60, 4 (April 2017), 86–93. https://doi.org/10.1145/3051088

Power analysis

- Ideally more than "In X, the authors used N samples and managed to get a publication, hence we use the same number".
- It's difficult or impossible for many microbiome applications.
- Try to get away without doing it.

Some resources

Brendan J. Kelly, Robert Gross, Kyle Bittinger, Scott Sherrill-Mix, James D. Lewis, Ronald G. Collman, Frederic D. Bushman, Hongzhe Li, *Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA*, Bioinformatics, Volume 31, Issue 15, 1 August 2015, Pages 2461–2468, https://doi.org/10.1093/bioinformatics/btv183

https://medium.com/brown-compbiocore/power-analyses-for-microbiome-studies-with-micropower-8ff28b36dfe3

Kers JG and Saccenti E (2022) *The Power of Microbiome Studies: Some Considerations on Which Alpha and Beta Metrics to Use and How to Report Results.* Front. Microbiol. 12:796025. doi:10.3389/fmicb.2021.796025

Ferdous, T., Jiang, L., Dinu, I. et al. *The rise to power of the microbiome: power and sample size calculation for microbiome studies.* Mucosal Immunol (2022). https://doi.org/10.1038/s41385-022-00548-1

Li Chen, *powmic: an R package for power assessment in microbiome case–control studies,* Bioinformatics, Volume 36, Issue 11, June 2020, Pages 3563–3565, https://doi.org/10.1093/bioinformatics/btaa197

To impute or not to impute?

- Imputation = making up data to make your algorithm/statistical test happy
- You may hear of it in other contexts
- Don't do it!

Issues to consider

- The stool it's not where cool things happen
- Diet (strongly) impacts gut microbiome
- Most GI disruptions lead to similar microbiome effects (e.g., more aerotolerant bacteria)
- Medication may impact gut microbiome (e.g., T2D)
- Temperature, pH, light, nutrients impact environmental communities

Animal "husbandry" matters

- Animals from different sources (even rooms) have different microbiota
- Co-housing, coprophagia, etc. need to be accounted for
- Huge variability among mammal-associated microbiota

Measurement biases

- Lysis biases
- 16S rRNA copy number
- PCR amplification biases
- Sequencing impacted by base composition
- Long read technologies prefer short fragments (and countermeasures focus on HMW DNA)

Sample processing matters

- Freezing kills bacteria
- Oxygen kills bacteria
- Time kills RNA
- Time changes compositions
- Host DNA contaminates samples
- Kit DNA impacts low-concentration samples

Use blank and positive controls!
(even if they cost you money)

Paired-end data

- For amplicon – ends should overlap

  read_len * 2 > amplicon_len
- For metagenomics

  fragment_len > ~4* read_len

  (useful for assembly and QC)

My controversial claim

- A well-designed study doesn't need fancy statistics!

Proving causality requires interventions

(irrespective of what your friendly statistician may say)

Resources

- Mallick, H., Ma, S., Franzosa, E.A. et al. Experimental design and quantitative analysis of microbial community multiomics. Genome Biol 18, 228 (2017). https://doi.org/10.1186/s13059-017-1359-z
- Shankar, J. Insights into study design and statistical analyses in translational microbiome studies. https://atm.amegroups.com/article/view/13582/html